Documento metodológico

Diseño Muestral y Ponderadores de la ENCC 2022/23

La Encuesta Nacional de Consumos Culturales (ENCC) de 2022/23 es una encuesta basada en una muestra probabilística, polietápica y estratificada, así como los relevamientos previos de 2017 y 2013. El universo de población cubierto por la misma son las personas de 13 y más años que residen en localidades urbanas de más de 30.000 habitantes de la República Argentina, según los datos del Censo Nacional de Viviendas del Año 2010 (CEN2010). A continuación, se detallan las distintas etapas de selección del muestreo de la ENCC-2022/23:

Etapa 1. Selección de localidades
Etapa 2. Selección de radios censales
Etapa 3. Selección de hogares
Etapa 4. Selección de una persona (entrevistado/a)

Las unidades de muestreo finales de la ENCC 2022/23 fueron una persona por hogar de la vivienda seleccionada de acuerdo con los procedimientos del diseño.

1. Características de las etapas de selección

1.1 Selección de localidades

La selección de localidades conservó las mismas características y listado del operativo anterior (ENCC 2017), con el propósito de reducir el error al momento de efectuar comparaciones con los resultados de esa edición. Uno de los principales criterios de selección de las unidades de muestreo en esta etapa fue el de estratificación según región geográfica, a través de un muestreo sistemático, ordenado geográficamente por localidades.

Cabe señalar que tanto la Ciudad Autónoma de Buenos Aires (CABA) como los departamentos (municipios) de la conurbación perteneciente al Gran Buenos Aires (GBA) conformaron regiones autorrepresentadas, debido a sus características poblacionales. Se seleccionaron, en cada región, las localidades con una probabilidad de selección proporcional al tamaño de la población (de acuerdo con los datos de población del CEN2010). Algunas localidades fueron seleccionadas con probabilidad igual a uno (unidades autorrepresentadas).

Así, la probabilidad de selección P_{ri} de una localidad *i* perteneciente a una región *r* es:

En caso de una localidad autoponderada $P_{ri} = 1$
Si no se trata de una localidad autoponderada $P_{ri} = n_r * (Pob_{ri} / Pob_r)$ donde

Pob_{ri} es la población del CEN2010 de la localidad i perteneciente a una región r
Pob_r es la población del CEN2010 de la localidad i perteneciente a una región r

El factor de expansión de una localidad *i* de una región *r* será $F_{ri} = 1 / P_{ri}$

1.2 Selección de radios censales

Dentro de cada localidad *i* seleccionada en una región *r*, se seleccionó una determinada cantidad de radios m_{ri} en forma sistemática con un intervalo entero. De este modo, la probabilidad de selección de un radio *j* en una localidad *i* de una región *r* será:

$P_{rij} = n_r * (m_{ri} / M_{ri})$ donde:

m_{ri} es la cantidad de radios a seleccionar en la localidad i de la región r
M_{ri} es la cantidad de radios en la localidad i de la región r

El factor de expansión de un radio j de una localidad i de una región r será:

$F_{rij} = 1 / P_{rij}$

1.3 Selección de hogares y personas a encuestar

En la última etapa se seleccionan, en cada radio previamente seleccionado, la vivienda, hogar y persona correspondiente a encuestar. Considerando las tasas de no respuesta obtenidas en las pruebas piloto, se decidió llevar a cabo una estrategia que combinó una selección aleatoria de viviendas, mediante un muestreo sistemático (con sobre-muestra para compensar la no respuesta); y una selección por cuotas de sexo y tramos de edad. Como es habitual, se consideró este tipo de selección como una selección aleatoria simple de los hogares del radio. Y se realizó luego una selección aleatoria simple de una persona en el hogar, dentro del conjunto de personas del hogar de 13 y más años. Por lo tanto, la probabilidad de selección del hogar *h* será:

$P_{rijh} = l_{rljh} / L_{rljh}$ donde:

l_{rij} es la cantidad de hogares a seleccionar en el radio *j* de la localidad *i* de la región *r*
L_{rij} es la cantidad de hogares en el radio *j* de la localidad *i* de la región *r* (registrados en la etapa de conteo, no se usa aquí CEN2010)

El factor de expansión de cada hogar será:

$F_{rijh} = 1 / P_{rijh}$

Y dentro de un hogar *h*, la probabilidad de selección de una persona mayor de 13 años es:

$1/K_{rijh}$, donde

K_{rijh} es la cantidad de personas mayores de 13 años del hogar h

El factor de expansión de las personas en cada hogar será:

$F_{rijh} = K_{rijh}$

2. Cálculo del valores de expansión

2.1. Factor de expansión de diseño

Con los factores calculados para cada etapa de selección, se calculó el factor de diseño, que surge como el producto de los factores de cada una de las etapas:

pondera = F1 • F2 • F3 • F4

2.2. Calibración de los factores de expansión

La técnica utilizada para mejorar las estimaciones en las encuestas, tanto probabilísticas como no probabilísticas, es la calibración de factores y se lleva a cabo a partir del uso de información auxiliar proveniente de una fuente externa, que puede o no estar en el marco de muestreo del diseño. En este sentido, la información auxiliar externa que se utilice dependerá del tipo de encuesta. Los criterios que se consideraron para el uso de información externa-auxiliar en este caso fueron:

1) La calidad de la información
2) El nivel de actualización de la información
3) La asociación o correlación de la variable auxiliar con la variable bajo estudio
4) La metodología de medición de la/s variable/s auxiliar/es
5) El nivel de alteración de los factores de diseño (factores originales) a causa de la introducción de la/s variable/s auxiliar/es en la calibración. En general, los métodos de calibración permiten acotar los ajustes a cierto rango, a saber: no disminuir el factor original a menos de su cuarta parte o no aumentarlo más de tres veces.

La elaboración de los ajustes-calibración de los factores siguió los siguientes pasos:

Dado los factores originales de expansión F_i.
Dada una serie de variables auxiliares x_1..., x_k cuantitativas
cuyos totales conocemos: t_{x1}...t_{xk}
Se trata de encontrar entonces nuevos factores de expansión d_i para que

$$t_{xj} = \sum_i d_i * x_{ji}$$

para j = 1, ..., k

En otras palabras: se trata de encontrar que con los nuevos factores se estime sin error los totales t_{xj} con la condición de hacer mínimo, con algún criterio, los cocientes Fi/di. Es decir, que los nuevos factores se alejen lo menos posible de los factores originales.

2.3 Implementación de la calibración en la ENCC-2022/23

Para mejorar las estimaciones y corregir posibles sesgos en el proceso de selección, se calibraron los factores de expansión utilizando la Encuesta Anual Urbana de Hogares de INDEC (EAUH2022) para obtener los totales marginales. La calibración de la ENCC 2022/23 se realizó con el paquete *survey* de R. En la ENCC 2022/23 las variables que intervinieron en la calibración fueron:

- Total de población por sexo
- Total de población por región
- Total de población por sexo y tramo de edad, por región
- Total de población ocupada por región
- Total de población de 25 y más años con educación secundaria incompleta, por región
- Total de población de 25 y más años con educación secundaria completa, por región.

Luego de efectuar la calibración, se ajustaron los ponderadores por región para tener en cuenta el crecimiento demográfico 2017-2022/23, procedimiento necesario para realizar comparaciones entre las ediciones de la ENCC.

3. Estimaciones e intervalos de confianza

Cuando se trabaja con estimaciones provenientes de una muestra probabilística, como la de la ENCC-2022, es necesario considerar el error muestral, una medida del cual se obtiene a través del denominado *Desvío Estándar*. A partir del cálculo del Desvío Estándar (DS), es posible puede construir un intervalo numérico que, con un determinado nivel de confianza o margen de error, contendrá al valor de la estimación correspondiente.

La teoría del muestreo indica que para muestras relativamente grandes, con un nivel de confianza del 95%, el intervalo comprendido entre la estimación menos 1,96 veces el desvío estándar y la estimación más 1,96 veces el estándar, contendrá al valor que se desea estimar. Por consiguiente, si se desea estimar un determinado total X, siendo el valor estimado para la base de datos de la ENCC un intervalo de confianza del 95%, en la estimación $\hat{X}$ será:

$$intervalo\ de\ confianza = \left(\hat{X} - 1.96\,DS\left(\hat{X}\right)\ ;\ \ \hat{X} + 1.96\,DS\left(\hat{X}\right)\right)$$

Este intervalo de confianza de la estimación contendrá, con una confianza del 95%, el valor verdadero del total X (en este caso).

Además, a partir del desvío estándar (DS) se puede calcular el coeficiente de variación (CV) de una estimación $\hat{X}$, CV ($\hat{X}$). Este coeficiente surge a partir del cálculo del cociente entre el desvío estándar (DS) de una estimación y esa estimación. Este resultado suele expresarse en porcentajes. Por tanto, el coeficiente de variación (CV) de una estimación $\hat{X}$ se define como:

$$\mathrm{CV}\ (\hat{X}) = 100 * \frac{DS\ (\hat{X})}{\hat{X}} \qquad (\text{suponiendo } \hat{X} \neq 0)$$

El coeficiente de variación otorga una idea de la precisión relativa de la estimación. Cuanto más pequeño sea el coeficiente de variación (CV) más precisa será la estimación. Si bien corresponde al usuario/a determinar si una estimación con cierto coeficiente de variación le es útil o no para su objetivo (según sea el grado de precisión que necesite), se advierte, en líneas generales, que estimaciones con coeficiente de variación superiores al 20% deben ser tratadas con precaución.